高校数据集成系统的ETL设计与实现

王晓虹[1],刘 莹[2],张艳凤[3]
(1. 辽宁石油化工大学 计算机与通信工程学院,辽宁 抚顺 113001;
2. 东北大学 信息科学与工程学院,辽宁 沈阳 110819;
3. 沈阳农业大学,辽宁 沈阳 110122)

摘 要:高校数字化校园建设是高校教育信息化发展的一项重要任务。由于高校普遍采用不同的数据库系统来管理学校的一切事务,致使信息存在大量的冗余、不一致,乃至"信息孤岛"的现象,从而严重制约了高校的信息化的建设。基于ETL的数据集成技术能够很好地解决这个难题。以高校集成数字校园平台建设为背景,提出了高校数据集成系统中基于ETL平台的建设方案。系统将ETL技术与传统的数据集成技术相结合,采用适配器技术和中间件技术,处理大批量的历史数据、实时处理小批量的变化的数据,全方位地满足用户对所有数据处理的需求,解决了异构数据集成和共享问题。

关键词:ETL;数据集成;实时抽取
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2011)07-0186-04

Design and Implementation of ETL Based on University Data Integration System

WANG Xiao-hong[1], LIU Ying[2], ZHANG Yan-feng[3]
(1. School of Information Engineering, Liaoning University of Petroleum & Chemical Technology, Fushun 113001, China;
2. College of Information Science and Engineering, Northeastern University, Shenyang 110819, China;
3. Shenyang Agricultural University, Shenyang 110112, China)

Abstract: Digital campus construction is an important task for the higher education information development. Due to the different database commonly used by universities to manage all matters of the school system, there are a lot of redundant and inconsistent information or even the information island. These have seriously hampered the university information construction. The data integration technology based on ETL can solve this problem. Taking the university integrated digital campus platform as the background, the construction program of the university data integration system based on ETL is proposed. The system integrates the ETL technology with traditional technology, and use adapter technology and middleware technology to handle large quantities of historical data and to deal with small quantities of data in real-time, which can meet customer demand of all the data processing, and solve the problem of heterogeneous data integration and sharing.

Key words: ETL; data integration; real-time extraction

0 引 言

高校数字化校园建设是目前高校教育信息化发展的一项重要任务。高校数字化校园建设的目的是将高校内原有的诸多彼此分离的业务子系统进行整合、集成,形成一个无缝平台,从而整体提高高校的教学质量、办公效率,整体提升高校的竞争能力。

本系统利用ETL和中间件技术构建一个高校数字校园框架,从各个应用系统抽取共享数据,并向各业务部门提供数据订阅,形成支撑其他应用的底层数据平台。在此基础上详细阐述了集成系统的数据整合的设计与实现。系统提供了一个适合高校目前状况、且能够快速实现数据共享的方案[1-3]。

1 高校数据集成系统的总体架构

本系统结合ETL和中间件技术,以分层次的组件方式对需要的数据从各异构数据源进行集成并向上提供统一的接口。系统架构图见图1。

高校数据集成系统分为三个模块:管理模块、代理模块和ETL模块。管理模块主要完成系统的控制、配

收稿日期:2010-12-08;**修回日期**:2011-03-08
基金项目:国家自然科学基金(青年基金)(61003003)
作者简介:王晓虹(1970-),女,副教授,研究方向为数据库和数据仓库、网络安全、嵌入式。

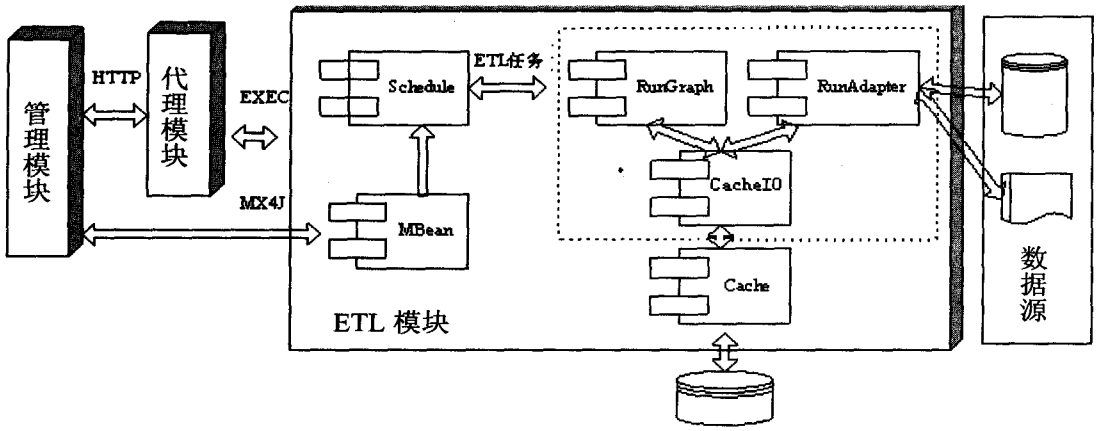

图1　数据集成系统架构图

置和监控等任务,它基于B/S结构运行在应用服务器上。代理模块主要完成ETL模块的启动和日志处理等任务,通过代理模块实现管理模块远程启动ETL模块的功能。ETL模块主要负责将分布的、异构数据源中的数据进行清洗、转换、集成,再加载到数据中心。每个代理模块可启动多个ETL模块,每个ETL模块可启动多个ETL任务[4-8]。

一个ETL模块由六个组件组成,分别是:包含MBean的JMX组件、Schedule组件、Cache组件、RunGraph组件、RunAdapter组件和CacheIO组件。JMX组件按照JMX规范实现管理功能,Mbean主要用于控制Schedule。Schedule组件主要完成任务调度,控制集成任务的生命周期。Cache组件主要完成从外部存储中读写数据及进行序列化支持。Cache增加了集成系统的安全性和可靠性,同时增加部署的灵活性。Cache未来可以替换成一个带有缓存传输功能的消息子系统或文件系统等。CacheIO组件调用Cache组件的接口,实现接口的转换。RunGraph组件用于控制运行一个Graph对象,RunAdapter组件用于运行一个Adapter对象。

ETL模块是进行ETL功能的子系统,每一个ETL模块可以包含多个Adapter组件和多个Graph组件完成一类任务的集成。每个Adapter组件直接访问一个被抽取或是一个被加载的资源。每个Graph包含一个或多个Node,每个Node是一个线程,完成一种相对独立的功能。Node和Adapter具备扩展能力,可以后期开发扩充。Graph与Adapter通过Cache组件完成数据的对接。

2　高校数据集成系统的ETL的关键技术

2.1　ETL模块的任务结构

每个ETL任务可以由数据抽取组件、数据转换组件和数据加载组件组合构成。执行方式为E、T、L三部分并发执行。ETL任务的内部结构图如图2所示。

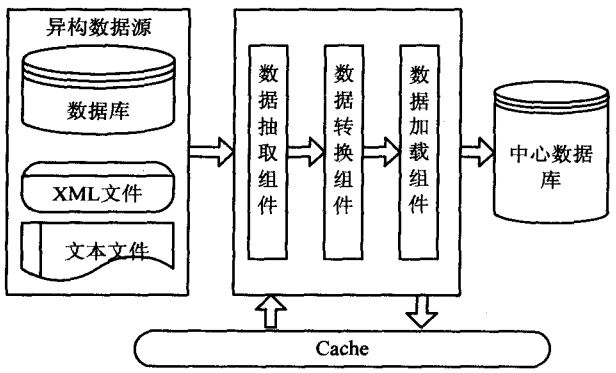

图2　ETL模块任务结构图

数据抽取组件从异构数据源(如数据库、XML文件、文本文件等)中定时抽取数据,数据抽取组件执行结束后,调用Cache API将数据存入Cache组件,Cache组件将二进制和大文本存入文件,根据数据抽取组件传入的数据生成OID和序列化数据,并写入在配置时生成的缓存表中。数据加载组件调用Cache API从Cache组件中读出数据。

2.2　ETL模块的处理流程

ETL是数据仓库实现过程中,将数据由数据源向数据仓库加载的主要过程。用户从数据源抽取出所需的数据,经过数据清洗转换,最终按照预先定义好的数据仓库模型,将数据加载到数据仓库中。ETL模块的处理流程如图3所示。根据不同的抽取策略进行参数、执行时间的初始化,之后进入后台守护模块实时监控。后台主要完成将业务数据按照抽取策略定时导入数据到临时数据库,处理后定时调用后台存储过程保存到数据中心库中。

2.3　集成数据抽取组件

2.3.1　主要功能

数据集成的基本需求是增量数据的监听和传输。根据高校集成系统目前的需求,数据抽取的类型分成两类:实时数据抽取和批量数据抽取。

(1) 实时数据抽取。

实时数据抽取采用基于消息技术的消息中间件方

式及分布式主流数据库之间的数据复制技术。实时数据特点是数据量不大,从一个字节到几百兆,同步或者异步传输。对于增量数据的清洗采用实时抽取转换加载。

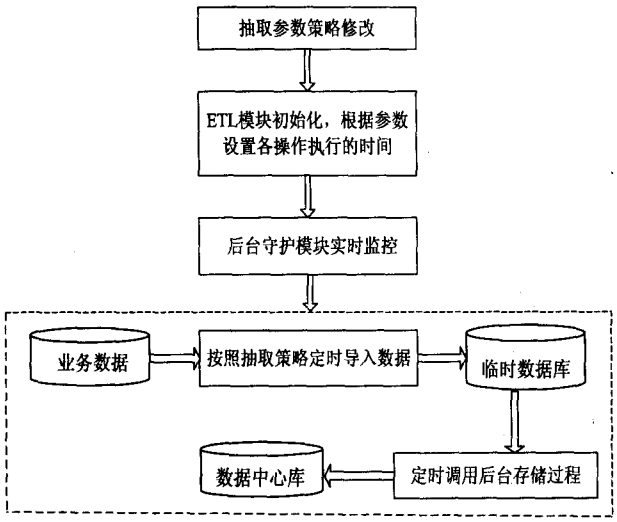

图 3 ETL 模块的处理流程

(2) 批量数据抽取。

批量数据抽取没有实时要求,但是量很大,达几百兆以上。批量数据抽取的实现主要是使用文件系统及数据处理工具等。批量加载和聚合运算依赖于 Adapter 的实现。

2.3.2 体系结构

数据抽取组件的体系结构图如图 4 所示。

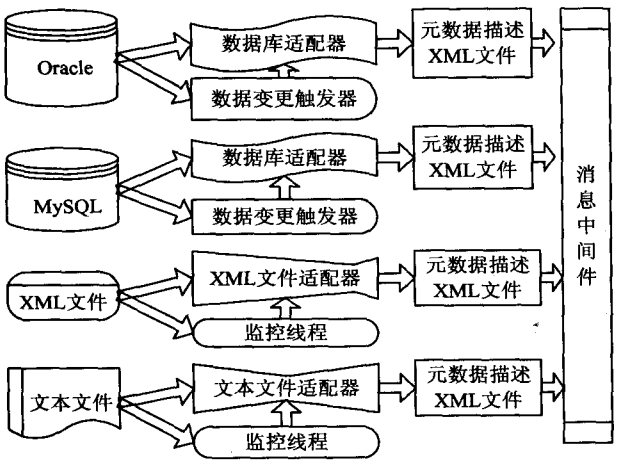

图 4 数据抽取组件的体系结构图

由于系统的集成针对不同平台、不同形式的源数据,本系统采用适配器来屏蔽各个数据源的异构性,向上提供一种一致的数据表达格式。适配器不但处理不同数据源的接入问题,同时还要完成从异构数据源抽取的数据转换成 XML 格式,形成元数据描述 XML 文件的功能。在系统中采用了数据库适配器、XML 适配器、文件适配器等来实现数据的抽取任务[9~11]。

2.3.3 实时抽取实现模块

高校数据集成系统既要处理历史数据,又要处理实时变更的业务数据。基于此特点,本系统实时抽取模块采用 CTF(Capture, Transformation, Flow)体系。该体系将 ETL 和传统的数据集成方法相结合,它既可以处理大批量的历史数据,又可以实时处理小批量的变化的数据,全方位地满足用户对所有数据处理的需求。

本系统主要采用以下模块实现实时抽取功能。

(1) 数据变更触发器。

在数据源的数据库中设立触发器,当数据发生写入、删除以及修改操作时,启动数据变更触发器,触发器设置数据变更事务,把发生变化的数据传送到适配器,适配器把处理好的数据通过消息中间件传送到数据转换器,由转换器处理完毕之后提交给中心数据库,若变更的数据已写入中心数据库中,则撤销此事务,否则继续执行此事务直到此事务提交完毕,最后关闭数据变更触发器。

(2) 文件变化监控线程。

利用专门的监控线程根据预先定义好的时间间隔轮询文件,对有变化的文件的数据进行抽取转换工作。

(3) 消息中间件。

基于消息的中间件技术有两种方式:消息传递方式和消息队列方式。

消息传递采用广播/预订方式。该通信模型提供了位置透明性功能。程序只需要简单地将消息以主题方式发送出去,由中间件来负责将消息传递给所有预订该主题的程序。消息中间件主要通过 agents 技术来实现 Publish-Subscribe 方式应用。当程序广播消息时,首先与一个代理进行连接,将消息传递给代理。代理负责路由消息给相应的程序,实现消息的动态路由功能。

消息队列方式允许程序无需直接建立起连接即可发送和接收消息。程序只须简单地将消息发送给消息队列,由消息队列负责消息的传递,对应用程序完全透明。消息队列采用异步方式,为信息提供了个安全的存储方式。

2.4 集成数据转换组件

根据集成系统的要求,系统以组件化的方式实现数据转换和加载。

集成数据转换是首先将一种数据库中定义的模型转化为另一种数据库中模型,再根据需要将源数据库中的数据转换到目的数据库中。对于整个转换分成模式转换和数据转换两部分。模式转换将源数据库中的数据字典转换成目的数据库中的数据字典;数据转换则按照模式转换的要求,从源数据库中数据经一系列转换产生目的数据库。

数据转换组件包括字段映射、数据过滤、数据清

洗、数据替换、数据计算、数据验证、数据加解密、数据合并、数据拆分等。这些组件可插拔,可任意组装,各组件之间通过数据总线共享数据。

数据清洗主要是针对源数据库中,对出现二义性、重复、不完整、违反业务或逻辑规则等问题的数据进行相应的清洗操作。在清洗之前需要进行数据质量分析,以找出存在问题的数据。数据清洗和纠错处理主要实现如下:对源数据的数据质量进行分析;建立清洗规则;根据清洗处理的结果进行纠错处理;差异比对和修复;比较任意两个数据库的两张表,可以指定比对规则;自动生成差异列表;根据比对的差异结果,自动修复目标表等[12]。

2.5 集成数据加载组件

集成数据加载将从数据源系统中抽取、转换后的数据加载到数据仓库系统中。数据加载策略要考虑加载周期及数据追加策略两方面的内容。根据高校教学、管理业务的实际情况,加载周期要综合考虑业务分析需求和系统加载的代价,对不同的子系统的数据采用不同的加载周期,但必须保存同一时间数据的完整性。

集成数据加载组件主要功能是解析元数据描述XML文件。XML文件解析模块将经数据转换层转换完成的元数据描述XML文件,通过对象-关系映射进行解析,最后将转换结果存入中心数据库中。

数据的追加策略根据数据的抽取策略和业务规则确定,本系统采用三种类型:直接追加、全部覆盖、更新追加。对于流水数据采用直接追加方式;如果抽取数据本身已包括数据的当前和所有历史状况,则对目标表采用全部覆盖方式;对于需要连续记录业务的状态变化,并用当前的最新状态同历史状态数据进行对比的情况可采用更新追加的方式。

3 结束语

根据高校数据集成系统的需求,构建了ETL系统模型和实现方案。该ETL系统基于适配器和中间件技术、具有多线程并发运行、增量抽取、易扩展、支持跨平台的功能。该系统可以很好地解决数据集成中多数据融合、数据不一致和数据同步更新等问题。

参考文献:

[1] 张上游,王　燕.数字化校园建设中数据分析的研究[J].西南民族大学学报,2009,35(4):871-873.

[2] 王晓虹,王国仁,于勇前,等.电信闭环决策支持系统的研究与实现[J].计算机应用研究,2008,25(4):1247-1249.

[3] 秦学勇,刘　栋.数据仓库的可扩展性研究与设计[J].计算机技术与发展,2009,19(5):65-67.

[4] 李　颖,郝克刚,葛　玮.基于电信数据仓库系统的ETL研究与设计[J].计算机应用与软件,2009,30(1):178-180.

[5] 许　力,牟晓光,马云存.并行ETL过程的研究与实现[J].计算机工程与应用,2009,45(13):170-172.

[6] 宋　杰,王大玲,鲍玉斌,等.一种元数据驱动的ETL方法的研究[J].小型微型计算机系统,2007,12(12):2167-2173.

[7] Panos V, Alkis S. Manolis G, et al. A generic and customizable framework for the design of ETL scenarios[J]. Information Systems,2005, 30(7):492-525.

[8] Alkis S, Dimitrios S, Malu C. Representation of conceptual ETL designs in natural language using Semantic Web technology[J]. Data & Knowledge Engineering,2010,69(1):96-115.

[9] 黄怀毅,杨路明.一种轻量级架构的ETL系统设计与实现[J].计算机技术与发展,2008,18(6):202-205.

[10] 戴　浩,杨　波.ETL中的数据增量抽取机制研究[J].计算机工程与设计,2009,30(23):5552-5555.

[11] Alkis S, Panos V, Timos S. State-Space Optimization of ETL Workflows[J]. IEEE Transactions on Knowledge & Data Engineering,2005, 17(10):1404-1419.

[12] 包从剑,李星毅,施化吉.可扩展和可交互的数据清洗系统[J].计算机技术与发展,2007,17(7):84-86.

(上接第185页)

(11):18-19.

[2] 吴英攀,于立新.基于层次化验证平台的存储器控制器功能验证[J].微电子学与计算机,2009(2):25-28.

[3] 王世好,王歆民,刘明业.嵌入式系统软硬件协同验证中软件验证方法[J].计算机研究与发展,2005,42:514-519.

[4] 马　宁,李　玲,田　泽,等.ARINC659总线协议芯片的仿真验证[J].计算机技术与发展,2010,20(1):205-206.

[5] 杨海波,田　泽,蔡叶芳,等.FC IP软核的仿真与验证[J].计算机技术与发展,2009,19(9):168-172.

[6] RapidIO Interconnect Specification Rev2.0,03[S].[s.l.]:Copyright RapidIO Trade Association,2008.

[7] 梁小虎.高速串行总线RapidIO与PCI Express协议分析比较[J].航空计算技术,2010(3):164-167.

[8] 黄振中,柴小丽.基于Vxworks的PCI-RapidIO桥驱动设计[J].计算机工程,2010(3):239-240.

[9] Serial RapidIO User Guide v5.1 UG503[M].[s.l.]:Xilinx,2008.

[10] Serial RapidIO User Guide v4.3 UG247[M].[s.l.]:Xilinx,2008.

[11] 郭　蒙,田　泽,蔡叶芳,等.1553B总线接口SoC验证平台的实现[J].航空计算技术,2008,38(6):99-101.

[12] 田　靖,田　泽.AFDX-ES SoC虚拟仿真平台的构建与应用[J].计算机技术与发展,2010,20(8):192-198.